jc511 U.S. PTO
09/537242
03/28/00

#3
1/17/00

Europäisches Patentamt	European Patent Office	Office européen des brevets

Bescheinigung | Certificate | Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. Patent application No. Demande de brevet n°

99480015.9

CERTIFIED COPY OF PRIORITY DOCUMENT

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

Arlette Fiedler

A. Fiedler

DEN HAAG, DEN

THIS PAGE BLANK (USPTO)

Europäisches Patentamt
European Patent Office
Office européen des brevets

Blatt 2 der Bescheinigung
Sheet 2 of the certificate
Page 2 de l'attestation

Anmeldung Nr.:
Application no.: 99480015.9
Demande n°:

Anmeldetag:
Date of filing: 30/03/99
Date de dépôt:

Anmelder:
Applicant(s):
Demandeur(s):
INTERNATIONAL BUSINESS MACHINES CORPORATION
Armonk, NY 10504
UNITED STATES OF AMERICA

Bezeichnung der Erfindung:
Title of the invention:
Titre de l'invention:
Router monitoring in a data transmission system utilizing a network dispatcher for a cluster of hosts

In Anspruch genommene Priorität(en) / Priority(ies) claimed / Priorité(s) revendiquée(s)

Staat: State: Pays:	Tag: Date: Date:	Aktenzeichen: File no. Numéro de dépôt:

Internationale Patentklassifikation:
International Patent classification:
Classification internationale des brevets:
/

Am Anmeldetag benannte Vertragstaaten:
Contracting states designated at date of filing: AT/BE/CH/CY/DE/DK/ES/FI/FR/GB/GR/IE/IT/LI/LU/MC/NL/PT/SE
Etats contractants désignés lors du depôt:

Bemerkungen:
Remarks:
Remarques:

THIS PAGE BLANK (USPTO)

ROUTER MONITORING IN A DATA TRANSMISSION SYSTEM UTILIZING A NETWORK DISPATCHER FOR A CLUSTER OF HOSTS

Technical field

The present invention deals with a new way for obtaining high availability and load balancing on default routers for IP host systems, and relates in particular to a router monitoring in such a system utilizing a network dispatcher for a cluster of hosts.

Background

Several types of digital networks operating with the packet switching technique in which data from different origins are chopped into fixed or variable length packets or datagrams have been installed throughout the world, which need to be interconnected (e.g. via so called Routers) to optimize the possibilities of organizing traffic between source hosts and target hosts located anywhere in the world. This is made possible by using so-called internetworking.

Internetwork (also referred to as internet) facilities use a set of networking protocols such as Transmission Control Protocol/internet Protocol (TCP/IP) developed to allow cooperating host computers to share resources across the internetwork. This is made possible by using so-called internetworking.

Internetwork (also referred to as internet) facilities use a set of networking protocols such as Transmission Control Protocol/Internet Protocol (TCP/IP) developed to allow cooperating host computers to share resources across the internetwork. TCP/IP is a set of data communication protocols that are referred to as internet protocol (IP) suite. Because TCP and

Printed:06-05-1999
EDOCID: <E1 9948001502>

IP are the best known, it has become common to use the term TCP/IP to refer to the whole protocol family. TCP and IP are two of the protocols in this suite. Other protocols of the suite are User Datagram Protocol (UDP), Address Resolution Protocol (ARP), Real Time Protocol (RTP) etc...

An internet may thus be a collection of heterogeneous and independent networks using TCP/IP, and connected together by routers. The administrative responsibilities for an internet (e.g. to assign IP addresses and domain names) can be within a single network (LAN) or distributed among multiple networks.

When a communication of data has to be established from a source host to a particular IP destination over an IP network, there is a number of methods to determine the first hop router of the network towards this destination. These include running (or snooping) dynamic routing protocol such as Routing Information Protocol (RIP) or Open Shortest Path First (OSPF) version, running an ICMP router discovery client or using a statically configured default route.

Running a dynamic routing protocol on every end-host may be infeasible for a number of reasons, including administrative overhead, processing overhead, security issues, or lack of a protocol implementation for some platforms. Neighbor or router discovery protocols may require active participation by all hosts on a network, leading to large timer values to reduce protocol overhead in face of large numbers of hosts. This can result in a significant delay in the detection of a lost (i.e., dead) neighbor, which may introduce unacceptably long «black hole» periods.

The use of a statically configured default route is quite popular, it minimizes configuration and processing overhead on the end-host and is supported by virtually every IP implementation. This mode of operation is likely to persist as

Dynamic Host Configuration Protocols (DHCP) are deployed, which typically provide configuration for an end-host IP address and default gateway. However, this creates a single point of failure. Loss of the default router results in a catastrophic event, isolating all end-hosts that are unable to detect any alternate path that may be available.

One solution to solve this problem is to allow hosts to appear to use a single router and to maintain connectivity even if the actual first hop router they are using fails. Multiple routers participate in this protocol and in concert create the illusion of a single virtual router. The protocol ensures that one and only one of the routers is forwarding packets on behalf of the virtual router. End hosts forward their packets to the virtual router. The router forwarding packets is known as the active router. A standby router is selected to replace the active router should it fail. The protocol provides a mechanism for determining active and standby routers, using the IP addresses on the participating routers. If an active router fails, a standby router can take over without a major interruption in the host's connectivity.

Another similar approach is the use of Virtual Router Redundancy Protocol (VRRP) designed to eliminate the single point of failure inherent in the static default routed environment. VRRP specifies an election protocol that dynamically assigns responsibility for a virtual router to one of the VRRP routers on a LAN. The VRRP router controlling the IP address(es) associated with a virtual router is called the Master, and forwards packets sent to these IP addresses. The election process provides dynamic fail-over in the forwarding responsibility should the Master become unavailable. Any of the virtual router's IP addresses on a LAN can then be used as the default first hop router by end-hosts. The advantage gained from using VRRP is a higher availability default path without requiring

BNSDOCID: <E1 9948001502>

configuration of dynamic routing or router discovery protocols on every end-host.

Unfortunately the two above solutions cannot provide load balancing for a given host's traffic because only the router that answered the ARP is used. Also, customers are reluctant to change their main router configuration to enable such a function.

It is why in IBM docket FR 9 99 008 filed as a European patent application, an IP source host is provided with a new layer between the IP layer and the network layer for selecting dynamically a router amongst a set of candidate default routers, thereby ensuring both load balancing and high availability.

Unfortunately, in case of a configuration with a network dispatcher used as a front end to a cluster of hosts, a host will always receive incoming packets from the network dispatcher in response to ARP requests, as opposed to packets from the candidate routers. Therefore, it is not possible to maintain the status of active candidate routers by resetting the age of an entry in the ARP table each time a packet is received from a matching network (MAC) address as in IBM docket FR 9 99 008 The only solution is to issue periodic ARP requests to candidate routers with the drawback that all the hosts have to monitor all the individual routers.

Summary of the invention

Accordingly, the object of the invention is to provide a specific device for monitoring all the candidate routers in a data transmission system wherein a cluster of hosts is associated to a network dispatcher receiving all the incoming flows from an IP network.

Another object of the invention is to achieve a method of determining the availability of candidate routers in a data transmission system wherein a cluster of hosts is associated to a network dispatcher receiving all the incoming flows from an IP network.

The invention relates therefore to a data transmission system for exchanging packetized data between any IP host amongst a cluster of IP hosts having each at least an IP layer and a network layer and a plurality of workstations by the intermediary of an IP network, wherein each IP host is connected to the IP network via a layer 2 network interfacing the IP network by a set of routers and by a network dispatcher in charge of receiving all incoming flows from the workstations and dispatching them amongst the cluster of hosts. Such a system comprises at least one monitoring device included in the cluster of hosts comprising means for monitoring the availability of the routers and means for broadcasting the router availability information to each host of the cluster of hosts via the network dispatcher.

Brief description of the drawings

The above and other objects, features and advantages of the invention will be better understood by reading the following more particular description of the invention in conjunction with the accompanying drawings wherein :

Fig. 1 represents schematically a data transmission system wherein a cluster of hosts incorporates a specific device for monitoring the availability of a set of routers according to the invention,

Fig. 2 is a flow chart of the method implemented in the invention for monitoring the availability of the routers.

Detailed description of the invention

In reference to Fig. 1, the invention is implemented in a data transmission system wherein a plurality of IP Hosts 10, 12, 14 transmit data to one or several workstations 16, 18, 20 via an IP network 22 by means of a layer 2 network such as a Local Area Network (LAN) 24. LAN 24 is interfacing IP network 22 by a set of input routers such as routers 26, 28. The IP packets are routed over the IP network via a plurality of routers (not shown) until output routers such as routers 30, 32 connected to workstations 16, 18, 20.

Instead of using a single default router to transmit data over the IP network, a technique described in IBM docket FR 9 99 008 consists for a host in using a new layer between the IP layer and the network layer, this additional layer being in charge of selecting one amongst a set of candidate routers such as routers 26 or 28 by running an algorithm based upon parameters defined in the packet which is transmitted.

In the present invention, it is assumed that the hosts 10, 12, 14 are grouped in a cluster associated with a network dispatcher interfacing LAN 24 with IP network 22. Such a network dispatcher (ND) is a solution to the problems of keeping the load evenly spread or balanced on a group of hosts (or servers). It acts as a dispatcher of connections from users who know a single IP address for a service, to the set of hosts 10, 12 and 14 which actually perform the work. Only the packets going from the users such as workstations 16, 18 and 20 pass through network dispatcher 34. The packets from an IP host to a workstation may go by other routes which need not include the network dispatcher 34, thereby reducing the load on a network dispatcher and allowing it to potentially stand in front of a larger number of hosts.

Since the cluster of hosts is seen by the users as the single address of the network dispatcher, it is therefore impossible to determine the availability of a router 26, 28 amongst the set of routers interfacing the IP network by only monitoring the data packets received from the IP network as mentioned above. The solution of the invention is therefore to add a router monitoring device (RM) 36 as a new member of the cluster of hosts. Instead of requiring each IP host to send ARP requests to each candidate router 26 or 28 in order to determine the availability of the latter, RM device 36 is in charge of sending periodically (the period can be short, e.g. 1 to 10 seconds in order to ensure the best service) a unicast ARP request to all the candidate routers, and then to inform all the IP hosts about the availability status of each router using a broadcast ARP response. Thus, this function is performed with a minimum traffic and the number of IP hosts (or servers) using the set of candidate routers can scale up without increased control traffic.

It must be noted that the function of such router monitoring device 36 may be integrated in one of the IP hosts. Furthermore, there can be several router monitoring devices or several IP hosts achieving this function.

Referring now to Fig. 2, the steps implemented in the invention are as follows. First, an ARP request (preferably a unicast request to all candidate routers) is sent to a router (step 40) on a periodical basis by router monitoring device 36. Then , it is checked (step 42) whether an answer is received by RM device 36 from the router. If so, this means that the router is available and an ARP response packet is sent by the RM device to all IP hosts (step 44). This response is preferably a MAC level broadcast indicating the IP address and the MAC address of the candidate router which has been requested as information indicating the availability of the

router. This response forces all the IP hosts to update their corresponding entry in the ARP table (step 46).

When no answer is received from the candidate router being requested (step 42), a test is made (step 48) to determine whether a decision factor is reached. For example, a router which fails to answer three times in a row can be declared unavailable. But, the decision factor could be a different one. Assuming the decision is reached, the RM device sends (step 50) an ARP response as a MAC level broadcast to all the IP hosts. This response indicates the IP address of the router and its MAC address set to a default value such as all zeroes as information indicating the unavailability of the router. This forces all the IP hosts to update their ARP table (step 52) by removing the ARP entry corresponding to the unavailable router after recognizing the invalid MAC address. Note that the entry can be updated with the invalid MAC address (e.g. all zeroes) rather than removing the entry.

Claims

1. Data transmission system for exchanging packetized data between any IP host amongst a cluster of IP hosts (10, 12, 14) having each at least an IP layer and a network layer and a plurality of workstations (16, 18, 20) by the intermediary of an IP network (22), wherein each IP host is connected to said IP network via a layer 2 network (24) interfacing said IP network by a set of routers (26, 28) and by a network dispatcher (34) in charge of receiving all incoming flows from said workstations and dispatching them amongst said cluster of hosts;

said system being characterized in that it comprises at least a monitoring device (36) included in said cluster of hosts comprising means for monitoring the availability of said routers and means for broadcasting the router availability information to each host of said cluster of hosts via said network dispatcher.

2. Data transmission system according to claim 1, wherein said at least monitoring device (36) is incorporated in one of said cluster of IP hosts (10, 12, 14).

3. Data transmission system according to claim 1 or 2, wherein said means for monitoring availability of said routers send periodically a unicast ARP request to all the candidate routers (26, 28).

4. Data transmission system according to claim 3, wherein said unicast ARP request to all the candidate routers (26, 28) is sent on a periodic basis comprised between 1 and 10 seconds.

5. Data transmission system according to any one of claims 1 to 4, wherein said means for broadcasting the router

BNSDOCID: <E1 9948001502>

availability send a MAC level broadcast indicating the IP address of the router being requested and an information on the availability of said router.

6. Data transmission system according to claim 5, wherein said information on the availability of the router is the MAC address of said router when this one has answered and is available.

7. Data transmission system according to claim 6, wherein said IP hosts (10, 12, 14) update their ARP table when receiving the MAC address of said router being requested.

8. Data transmission system according to claim 5, wherein said information on the availability of the router is a default value like all zeroes of the MAC address of said router when said router is considered unavailable.

9. Data transmission system according to claim 8, wherein said IP hosts (10, 12, 14) update their ARP table by removing the corresponding entry or writing said default value when said router is considered unavailable.

10. Data transmission system according to claim 8 or 9 wherein said router being requested is considered unavailable when it has not answered three monitoring requests in a row from said router monitoring device (36).

11. Method of determining the availability of candidate routers in a data transmission system for exchanging packetized data between any IP host amongst a cluster of IP hosts (10, 12, 14) having each at least an IP layer and a network layer and a plurality of workstations (16, 18, 20) by the intermediary of an IP network (22), wherein each IP host is connected to said IP network via

a layer 2 network (24) interfacing said IP network by a set of routers (26, 28) and by a network dispatcher (34) in charge of receiving all incoming flows from said work-stations and dispatching them amongst said cluster of hosts, said method being characterized in that a unicast ARP request is sent periodically to all candidate routers and a MAC level broadcast is then transmitted to all IP hosts for them to update their ARP table with the router information on the availability.

12. Method according to claim 11, wherein said information on the availability of the router is the MAC address of said router when this one has answered and is available.

13. Method according to claim 12, wherein said IP hosts (10, 12, 14) update their ARP table when receiving the MAC address of said router being requested.

14. Method according to claim 11, wherein said information on the availability of the router is a default value like all zeroes of the MAC address of said router when said router is considered unavailable.

15. Method according to claim 14, wherein said IP hosts (10, 12, 14) update their ARP table by removing the corresponding entry or writing said default value when said router is considered unavailable.

16. Method according to claim 14 or 15, wherein said router being requested is considered unavailable when it has not answered to three monitoring requests in a row from said router monitoring device (36).

THIS PAGE BLANK (USPTO)

ROUTER MONITORING IN A DATA TRANSMISSION SYSTEM UTILIZING A NETWORK DISPATCHER FOR A CLUSTER OF HOSTS

Abstract

Data transmission system for exchanging packetized data between any IP host amongst a cluster of IP hosts (10, 12, 14) having each at least an IP layer and a network layer and a plurality of workstations (16, 18, 20) by the intermediary of an IP network (22), wherein each IP host is connected to the IP network via a layer 2 network (24) such a LAN interfacing the IP network by a set of routers (26, 28) and by a network dispatcher (34) in charge of receiving all incoming flows from the workstations and dispatching them amongst the cluster of hosts. The system comprises at least a monitoring device (36) included in the cluster of hosts comprising means for monitoring the availability of the candidate routers and means for broadcasting the router availability information to each host of the cluster of hosts via the network dispatcher (34).

Fig. 1

THIS PAGE BLANK (USPTO)

FR 9 99 018
Lamberton et al
1/2

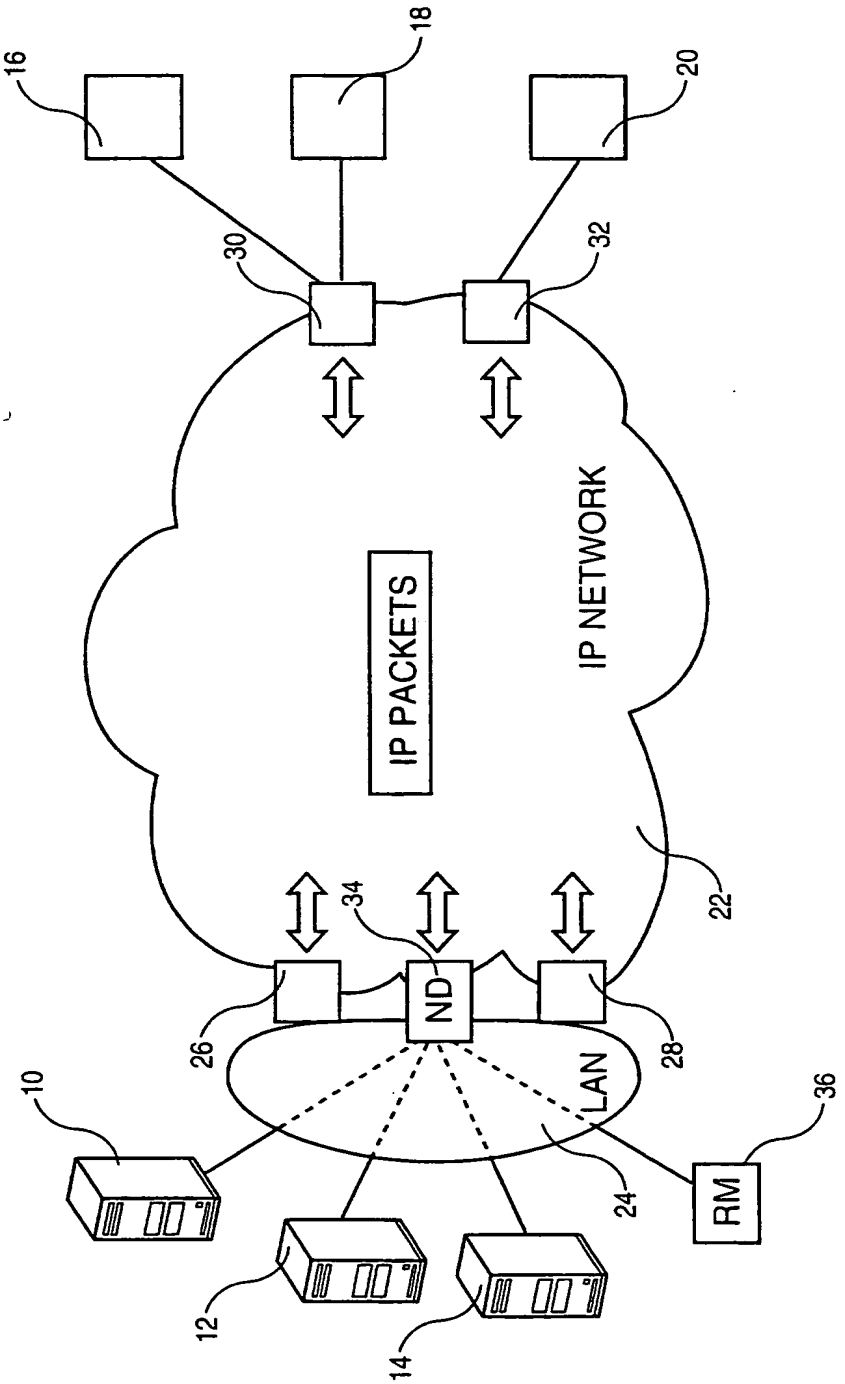

FIG. 1

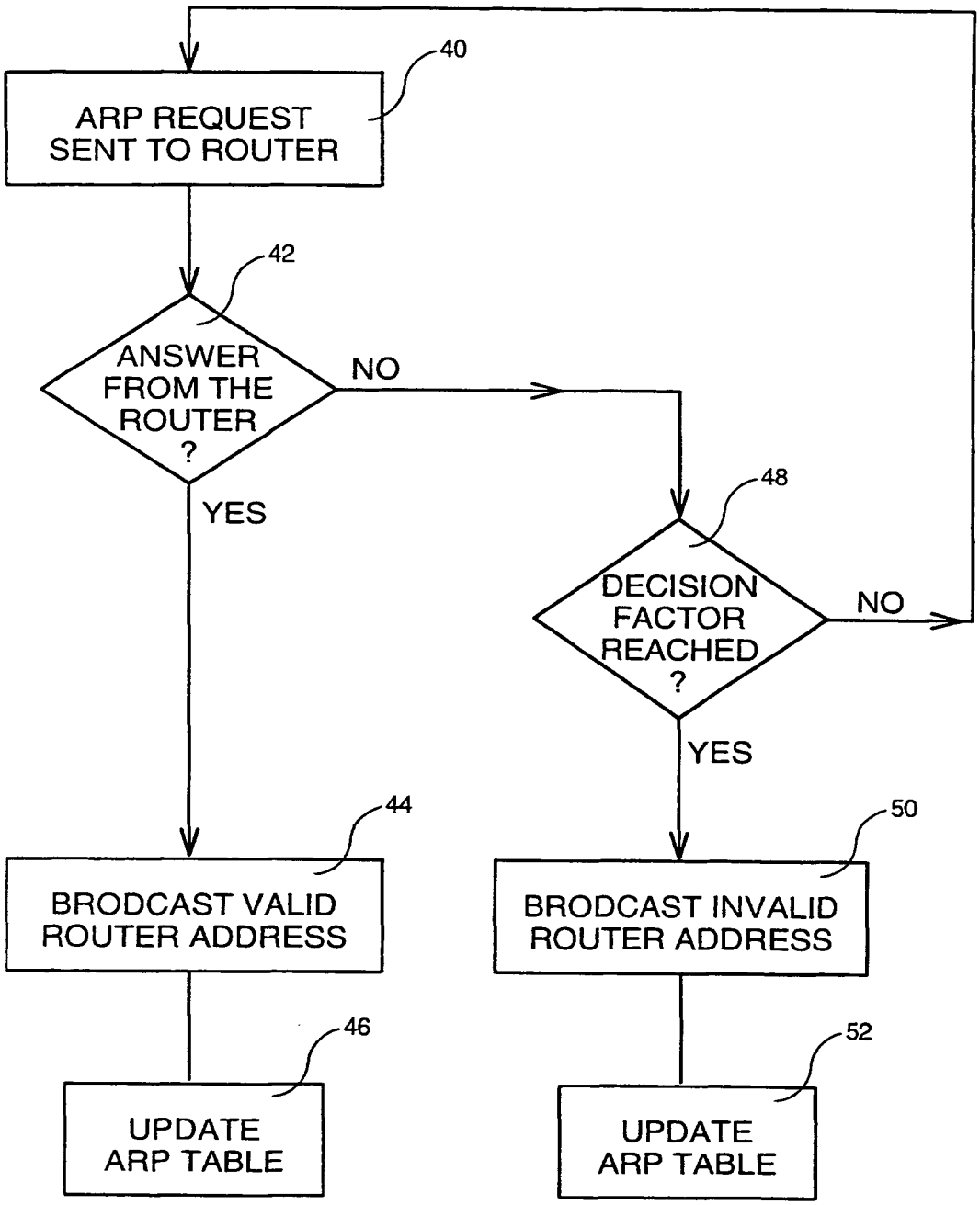
40
ARP REQUEST SENT TO ROUTER
42
ANSWER FROM THE ROUTER ?
NO
YES
48
DECISION FACTOR REACHED ?
NO
YES
44
BRODCAST VALID ROUTER ADDRESS
50
BRODCAST INVALID ROUTER ADDRESS
46
UPDATE ARP TABLE
52
UPDATE ARP TABLE

FIG. 2